DOCUMENT RESUME

ED 420 692 TM 028 367

AUTHOR	Johanson, George; Motlomelo, Samuel
TITLE	An Item Format Continuum for Classroom Assessment.
PUB DATE	1998-04-00
NOTE	10p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Educational Assessment; Essay Tests; *Measurement Techniques; *Objective Tests; Test Construction; *Test Format; Test Items; Test Validity

ABSTRACT

Many textbooks in educational measurement and classroom assessment have chapters devoted to specific item formats. There may be attempts to relate one item format to another, but the chapters and item formats are largely seem as distinct entities with only loose and uncertain connections. This paper synthesizes these discussions. An item format continuum is suggested. This continuum closely resembles the work of T. Rocklin (1992), R. Bennett (1993), and R. Snow (1993). There are seven broad categories of test items: (1) dialogue-oral; (2) performance; (3) project; (4) essay; (5) short answer; (6) multiple choice; and (7) true-false. Test validity has not been overlooked as a characteristic, but is simply not related to this continuum. Selecting an appropriate item format for a valid assessment of an instructional unit requires that the teacher match his or her objectives with item format characteristics. An appendix discusses the characteristics of the item formats in the continuum. (Contains 22 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
George Johanson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

An Item Format Continuum for Classroom Assessment

George Johanson and Samuel Motlomelo
College of Education
Ohio University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

A Paper Presented at the 1998 Annual Meeting of the
American Educational Research Association, San Diego, CA

ED 420 692

Many textbooks in educational measurement and classroom assessment have chapters devoted to specific item formats. That is, it is not uncommon to find a chapter concerning, say, the essay format. In such a chapter, the strengths and weaknesses of the particular format are often noted along with guidelines for the construction and use of such items. There may be attempts on the part of the author(s) to relate some of the characteristics of one format to another, but the chapters and the item formats, are largely seen as distinct entities with only loose and uncertain connections. We propose a synthesis of these chapters.

A number of authors have presented schemes for use in viewing and comparing the various item formats used in classroom assessment today (Bennett, 1993; Snow, 1993). In this paper, we suggest a similar item format continuum and, most importantly, discuss some of the characteristics of this continuum (see Appendix: Characteristics of the Continuum). The related literature surrounding these characteristics will also be briefly discussed. Our greater purposes are to first point out most clearly that every item format has strengths and weaknesses and that these are sometimes, happily, complementary; there is simply no single item format that is superior for all educational or classroom purposes on all occasions at all levels with all students. Second, a broader recognition of these item format characteristics may serve to govern at least the amplitude of the swings of the educational assessment pendulum. Third, the item format continuum we propose may provide a useful and integrating instructional device for those who teach and learn about educational assessment.

Item Format Continua in the Recent Literature

Rocklin (1992) did a multidimensional scaling of college student perceptions of similarity between pairs of (8) test item formats. The first dimension of the two dimensional MDS solution had the essay format to the left of the short-answer format which was, in turn, left of the multiple-choice format. The true-false item format was to the far right on this first dimension which was interpreted, in part, as separating supply-type items from selection-type items.

Bennett (1993) presented a scheme for categorizing item types in which the "organization reflects a hypothetical gradation in the constraint exerted on the nature and extent of the response" (p. 2). That is, the framework reflects the extent of the (student) construction of the response. The item format with the least construction is at the top and that with the most (but not necessarily most complex) construction is at the bottom of the following list:

- Multiple choice (choose the correct response from a small number of options)
- Selection/identification (choosing from a large number of options)
- Reordering/rearrangement (choosing an arrangement--- perhaps a logical ordering)
- Substitution/correction (replacement is the task, not choice)

TM028367

- Completion (a specific stimulus to supply a response, e.g., fill-in-the-blank)
- Construction (the construction of a complete response, e.g., an essay test item)
- Presentation (physical presentation or performance)

Note the similarity between the first dimension of Rocklin's (1992) similarity scaling (derived from student ratings) of item formats and Bennett's (1993) scheme which is based on the extent of student response construction.

Snow (1993) also presented a continuum of eight constructed-response test formats ranging from least response construction at the top to most at the bottom (p. 48):

- Multiple choice
- Multiple choice with intervening construction
- Simple completion/cloze procedure
- Short answer essay/complex completion
- Problem exercise
- Teach-back procedure
- Long essay/demonstration/project
- Collection of above over time, portfolios, and so on

This is similar to the preceding structures.

The Proposed Item Format Continuum

Our suggestion for an item format continuum closely resembles the efforts previously discussed. We have also not attempted to include all possible item formats, but have selected those we consider most often used and those most useful for our purposes. The proposed continuum includes seven broad categories of test item format and is illustrated in Figure 1.

<insert Figure 1. about here>

Missing item formats such as the mathematical problem format, portfolios, matching items, the cloze format, multiple-true-false items, alternate choice items, and so on may be located, at least approximately (and not without debate) along the continuum once the continuum characteristics are observed. More generally, note that we also do not discuss some of the broader, yet related formatting issues such as computer administration of test items, computer simulations, computer adaptive testing, testlets, and other topics.

Characteristics of the continuum of item formats are discussed briefly in the appendix and are listed in Figure 2.

<insert Figure 2. about here>

Discussion

The first two purposes for our work are somewhat similar and reflect our belief that diversity of method in educational assessment is desirable. Interestingly, this may be even more important to a teacher who has focused instruction on the highest cognitive skills or the most complex understandings (Feltovich, Spiro, & Coulson, 1993):

For complex material, in both testing and instruction, it seems prudent not to do anything

> one way. Singular approaches are likely to be detrimental because they: (a) do not provide a wide enough "lens" on the numerous aspects of the material to be taught or understood, (b) are likely to miss the interconnectedness of the target material with other related material, and (c) reinforce a misleading orientation toward complex material, by suggesting that it is simpler than it really is. (p. 210)

An analogy often used in courses in methods of teaching and educational measurement is the importance of a teacher having a variety of 'tools' in his or her instructional and assessment toolkits. While most textbooks discuss different item formats, many do this in very separate chapters without an integrating discussion of substance. Students in such courses may have an opportunity to learn many of the measurement properties only within particular formats. As a consequence, students may also be less aware of the interrelationships among item formats and feel less inclined towards a desired level of diversity of method. This brings us to our third purpose.

Item format characteristic lists tend to invite reader contributions and debate. Such debate is highly desirable in an instructional setting for learning about classroom assessment. It can even be instructive to note characteristics that do not fit well within the proposed item format continuum. Our perspective (and perhaps one of our main points overall) is simply that test *validity*, conspicuous by its absence from the list of characteristics, was not overlooked as a characteristic, but is simply not related to this continuum. A valid classroom assessment must accurately reflect the objectives of instruction and evaluation and these will likely vary both within and among units of instruction. In short, selecting an appropriate item format for a valid assessment of an instructional unit requires that the teacher match his or her objectives with item format characteristics.

References

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Elliott Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum Associates.

Feltovich, P., Spiro, R., & Coulson, R. (1993). Learning, teaching, and testing for complex conceptual understanding. In N. Frederiksen, R. Mislevy, & I. Bejar (Eds.), Test theory for a new generation of tests (pp. 181-217). Hillsdale, NJ: Erlbaum.

Rocklin, T. (1992). A multidimensional scaling study of college student's perceptions of test item formats. Applied Measurement in Education, 5(2), 123-136.

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Elliott Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 45-60). Hillsdale, NJ: Lawrence Erlbaum Associates.

Figure 1. A suggested item format continuum for educational assessment.

Dialogue-Oral	Performance	Project	Essay	Short-Answer	Multiple-Choice	True-False
from a Socratic Dialogue to an Oral Exam	this might well include portfolios	could be both in and out of class	brief & extended response	constructed response such as fill-in-the-blanks	complex M-C and matching or multiple T-F	also alternate choice and corrections

Figure 2. Selected characteristics of the item format continuum.

Dialogue-Oral	Performance	Project	Essay	Short-Answer	Multiple-Choice	True-False
More realistic ('authentic')						Less realistic
Better for higher order cognitive skills						Better for lower order cognitive skills
Student provides structure						Teacher provides structure
Subjective (less reliable) scoring						Objective (more reliable) scoring
Larger learning component						Smaller learning component
More diagnostic (formative)						Less diagnostic (summative)
Narrower content coverage						Broader content coverage
Guessing less of a factor						Guessing more of a factor
Faking is more of a problem						Faking is less of a problem
Cost (scoring) at the end						Cost (construction) in the beginning
Focus on the Process & Product						Focus on the Product
Less instructional sensitivity						More instructional sensitivity
Discovery methods of learning are preferred						Drill and practice are the rule
Novel problems or applications on a test						Strict alignment of teaching and testing
More cognitive learning theories						More behavioral learning theories
Better suited for small-scale applications						Suited for large-scale or small-scale applications

An Item Format Continuum for Classroom Assessment
Appendix: Characteristics of the Continuum

The purpose of this appendix is to both briefly discuss the characteristics mentioned in our paper and to identify a small portion of the relevant literature regarding these characteristics. It must be noted that the characteristics we discuss do not form exhaustive list nor is the literature entirely consistent. This appendix is intended more as a point of departure for further discussion than as a definitive destination. A reasonable context for the following comments would be a classroom assessment with a fixed time period of, say, 40 minutes.

Dialogue-Oral	Performance	Project	Essay	Short-Answer	Multiple-Choice	True-False

More realistic ('authentic')----------Less realistic

Paper-and-pencil tests in general, and multiple-choice and true-false items in particular, will always be more artificial or less realistic when compared to actual performances or hands-on assessment activities. Boodoo (1993) stated that performance assessments promise authentic and direct appraisals of educational competence. She further suggested that authentic assessments aim to capture a richer array of students' knowledge and skill than is possible with multiple-choice tests. Linn & Gronlund (1995) note that the multiple-choice item may measure whether the student knows or understands what to do when faced with a problem situation, but it cannot determine how the student actually will perform in that situation.

Better for higher order cognitive skills----------Better for lower order cognitive skills

Bracey (1993) noted that teachers reported that multiple-choice questions tend to contain elements that measure trivial and contrived materials. He further notes that multiple-choice emphasizes 'factoids' and tiny well-structured problems. Kon & Martin-Kniep (1992) showed that performance tests offer an important alternative to multiple-choice tests. These authors suggest that by offering a wider range of test formats, students do get an excellent opportunity to show what they know and what they can do. They note that this is particularly true for the assessment of higher-order thinking skills, for which the performance tests seem to be particularly well-suited. In comparing performance tests with objective tests, Oosterhof (1994) points out that performance tests directly measure higher cognitive skills whereas objective tests are not able to measure high order skills directly.

Pollack (1990) argues that free-response offers students an opportunity to show what they can do, rather than what they can recognize. Recognition is a process which may well involve a lesser cognitive skill than recollection. It is often difficult to judge if a student gets a multiple-choice item correct as a result of recall or by merely recognizing the most appropriate choice. Referring to objectively scored item formats, Hanna (1993) writes "...these item types tend to be more useful for measuring examinee command of lower-level learning than for assessing their use of higher mental processes." (p. 135).

When comparing multiple true-false items to multiple-choice items, Downing et al. (1995) state that "Test developers may find that the MCQ (multiple-choice question) remains the most appropriate for measuring the so-called higher levels of the cognitive taxonomy." (p. 195).

Student provides structure ----------Teacher provides structure

In essay, oral, or performance examinations (and most projects and reports) students have both the opportunity and obligation to present their ideas using their own organizational skills or structure. Such activities usually require the preparation of lengthy written responses and the performance of complex skills. On the other hand selected response assessment both allows and requires teachers to provide the structure (Stiggins, 1994). We might note as well that a student is required to be more cognitively active where a task is provided with less structure.

Subjective (less reliable) scoring----------Objective (more reliable) scoring

Multiple-choice and true-false tests are more consistently scored than performance or essay tests. Hanna (1993) declares simply "Performance measures tend to be much less reliable than objective tests." (p. 249). In particular, given a scoring scheme (typically, a rubric or model for an essay examination, a rating form for a performance or oral assessment, or an answer key for an objective test) the inter-rater reliability will be nearly perfect for only the objective formats. Frary (1985) indicated that scoring errors might be more common in free-response items than in multiple-choice items. Objective items, in fact, are often scored using computers. Essays may have a number of scoring difficulties and, in many cases, even experts may not agree on scoring the examination (Gronlund & Linn, 1990).

Given a fixed testing time, the reliability of objective tests is also enhanced simply by having a greater number of items on the test. Bridgeman & Lewis (1994) note that "Because of measurement error created by subjective scoring and by the relatively narrow coverage of the content domain, essay tests may be substantially less reliable than multiple-choice tests in the general subject area." (p. 37). With respect to oral examinations, Hanna (1993) states "Compared with written essay or objective examinations, oral tests...provide less reliable results." (p. 224).

Larger learning component----------Smaller learning component

Performance oriented formats afford students an opportunity to analyze and synthesize information in their own way, using their own experiences. Performance test formats allow students to use problem-solving skills and high level thinking and reasoning (Wilson et al., 1974). In fact, many assessment formats (projects, performances, essays, and especially dialogues) are often used for instructional, as opposed to assessment, purposes. By contrast, the learning value of typical objective item formats ranges from negligible to negative.

If you accept the argument that involvement in activities that require a higher level of cognitive skill tends to promote more learning than involvement in less demanding activities, (such as simple recall or recognition), then the prior item characteristic concerning cognitive levels supports the current contention. Wilson et al. (1974) criticize objective formats for imposing upon the student the task of selecting one correct answer among two or more options or of just furnishing a word, a phrase, or possibly a sentence to complete the answer sought by the examiner. Nitko (1996) states "Since performance assessments are very close to the ultimate learning targets of schooling, they may be used as instructional tools." (p. 108).

More diagnostic (formative)----------Less diagnostic (summative)

Since essays, performance assessments, projects, and portfolios require students to express themselves in their own words or create a tangible product, it is a straight forward undertaking to identify areas of weakness or misconceptions. It may not be nearly as easy for a teacher to identify these same specific weaknesses or the nature of a misunderstanding of

students when the student has only selected a response from among those that the teacher has offered.

In discussing informal oral assessment techniques, Nitko (1996) says "These questions should encourage students to think about the material and to reveal their understandings, including misconceptions. This will help you guide your teaching". (p. 104).

Narrower content coverage----------Broader content coverage

Discussing multiple-choice and essay formats, Bridgeman & Lewis (1994) commented that "The two types of tests differ in their coverage of the content domain. Essay examinations usually require an in-depth understanding of a few content areas while multiple-choice examinations survey a broader range of topics." (p. 37). Green (1979) showed that an advantage of the multiple-choice format over the performance format is that multiple-choice allows a more efficient sampling of course content per unit time period (she also points out that the advantage may be offset by lowered reliability of multiple-choice due to guessing). Boodoo (1993) indicated that one reason for the popularity of the multiple-choice format is that it can assess a wide range of information in a time-efficient manner with acceptable reliability. Certainly, the greater depth of understanding tapped by essays, performance assessments, and oral examinations requires more time and (given a fixed amount of time) necessarily narrows or diminishes the content coverage possible using these item types.

Guessing less of a factor----------Guessing more of a factor

Guessing is more of a factor in objective testing than it is with performance formats according to Frisbie & Becker (1990). Oosterhof (1994) and many others note that multiple-choice items are susceptible to guessing, but says the probability of answering many items correctly as a result of guessing alone is very small. Of course, it is more difficult to guess successfully on supply-type item formats.

Faking is more of a problem----------Faking is less of a problem

Faking, on the other hand, can be much more of a problem in essay and performance formats than it would be in objective tests particularly if faking is defined to include a social desirability response. That is, test takers may well be aware of the desires of examiners and be tempted to respond accordingly. Certainly, faking 'smart' on objective tests tends to occur infrequently! Hopkins & Antes (1989) support these contentions about both guessing and faking.

Cost (scoring) at the end----------Cost (construction) in the beginning

Bridgeman & Lewis (1994) state that "Essay examinations assess productive and organizational skills that cannot be measured with multiple-choice questions, but they require time-consuming and expensive scoring sessions that can be run only with trained experts in the subject area of the examination. On the other hand, multiple-choice tests are easy to score with machines." (p. 37). There are different opinions on the issue of whether it is easier to construct items with an essay format or an objective format; good essays may well take nearly the same time and effort to construct that good multiple-choice items take. Others would support the view that it takes more time to construct tests with objective formats than essay formats (e.g., Carey, 1994). In any event, essays are never quick and easy to score. Payne (1992) noted that "...the scoring of essay items and tests is among the most time-consuming and frustrating tasks associated with conscientious classroom measurement" (p. 178). While most would agree that it takes time and effort to construct good objective tests, it is also true that such tests can be scored by machines or quickly and accurately by anyone who has been provided with an answer key.

Focus on the Process & Product----------Focus on the Product

McDaniel (1994) has noted that "... analytic examination of artwork or essay provides insights about how well various components of the production process have been handled" (p. 183). The focus of essays and portfolios is on both process and product (Oosterhof, 1994; Stiggins, 1994). Objective formats are often exclusively concerned with the product or outcome; it is difficult to 'show-your-work' on a multiple-choice examination. Performance assessments, in particular, attempt to provide more direct and realistic measures of skills and processes than objective tests. Paper-and-pencil testing often excludes access to the process; a teacher must observe a performance to have knowledge of the process.

Less instructional sensitivity----------More instructional sensitivity

Instructional sensitivity is a measure of the extent to which students gain skills from instruction. Gronlund (1988) says that instructionally sensitive items will be "...answered correctly by a larger number of students after instruction than before instruction" (p. 109). He uses an index of sensitivity to instructional effects as part of an item analysis for criterion-referenced tests. Hanna (1993) claims that instructional sensitivity is greater for items at the lower end of the cognitive hierarchy and less for items that assess higher order thinking skills.

Discovery methods of learning are preferred----------Drill and practice are the rule

If students are to develop (or supply) their own responses and products based on their understandings as in essay examinations and performance assessments, then instruction must encourage and give practice in this creative thinking. Conversely, if there is a specific set of materials to be mastered (e.g., multiplication facts or spelling words), then a more reasonable approach would be to provide opportunities for this very structured learning.

Novel problems or applications on a test----------Strict alignment of teaching and testing

Nitko (1996) states this quite clearly when he says "...you must use novel materials to assess higher-order thinking" (p. 177). However, if you want to convey factual information, for example, a list of spelling words, then you might first teach the correct spelling of the words and then test the student's ability to spell precisely those words on the list.

More cognitive learning theories----------More behavioral learning theories

Shepard (1991, p. 9) found that: "...approximately half of all measurement specialists operate from implicit learning theories that encourage close alignment of tests with curriculum and judicious teaching of tested content." Her conclusion was that "These beliefs, associated with criterion-referenced testing, derive from behaviorist learning theory...". By way of contrast, certainly much of the movement towards performance assessment is being driven by the newer cognitive learning theories.

Better suited for small-scale applications-------Suited for large-scale or small-scale applications

Performance oriented examinations have two major limitations which tend to discourage their use in large scale assessments. First, performance assessments are focused or small-samples of a much larger content domain. Second, they are neither easily nor inexpensively scored and tend to have low reliability. Objective tests have neither of these limitations and have been used successfully with both small and large groups for a number of years. Objective formats are commonly used in standard large-scale examinations such as the SAT and military entrance examinations (Stiggins, 1994; Linn & Gronlund, 1995).

References

Boodoo, G. M. (1993). Performance assessments or multiple choice. Educational Horizons, 72 (1), 50-56.

Bracey, G. W. (1993). Assessing the new assessments. Principal, 72(3), 34-36.

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. Journal of Educational Measurement, 31(1), 37-50.

Carey, L. M. (1994). Measuring and evaluating school learning (2nd ed.). Boston, MA: Allyn and Bacon.

Downing, S. M., Baranowski, R. A., Grosso, L. J., & Norcini, J. J. (1995). Item type and cognitive ability measured: The validity evidence for multiple true-false items in medical specialty certification. Applied Measurement in Education, 8(2), 187-197.

Frary, R. B. (1985). Multiple-choice versus free response: A simulation study. Journal of Educational Measurement, 22(1), 21-31.

Frisbie, D. A., & Becker, D. F. (1990). An analysis of textbook advice about true-false tests. Applied Measurement in Education, 4(1), 67-83.

Green, K. (1979). Multiple choice and true-false: Reliability and validity compared. Journal of Experimental Education, 48(1), 42-44.

Gronlund, N. E. (1988). How to construct achievement tests (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Gronlund, N. E., & Linn, R. L. (1990). Measurement and evaluation in teaching (6th ed.). New York: Macmillan.

Hanna, G. S. (1993). Better teaching through better measurement. Fort Worth, TX: Harcourt Brace Jovanovich.

Hopkins, C. D., & Antes, R. L. (1989). Classroom testing construction (2nd ed.). Itasca, IL: F. E. Peacock Publishers, Inc.

Kon, J .H., & Martin-Kniep, G. O. (1992). Students' geographic knowledge and skills in different kinds of tests: Multiple-choice versus performance assessment. Social Education, 56 (2), 95-98.

Linn, R. L., & Gronlund, N. E. (1995). Measurement and assessment in teaching (7th ed.). Englewood Cliffs, NJ: Merrill.

McDaniel, E. (1994). Understanding educational measurement. Madison, WI: Brown and Benchmark.

Nitko, A. J. (1996). Educational assessment of students (2nd ed.). Englewood Cliffs, NJ: Merrill.

Oosterhof, A. (1994). Classroom applications of educational measurement (2nd ed.). New York, NY: Merrill.

Payne, D. A. (1992). Measuring and evaluating educational outcomes. New York: Merrill.

Pollack, J. M. (1990). Some issues in free response testing (ERIC Document Reproduction Service No. ED 322 200).

Shepard, L. A. (1991). Psychometricians' beliefs about learning. Educational Researcher, 20(6), 2-16.

Stiggins, R. J. (1994). Student-centered classroom assessment. New York: Merrill.

Wilson, J. A. R., Robeck, M. C., & Michael, W. B. (1974). Psychological foundation of learning and teaching. New York: McGraw-Hill.